A UK perspective on responsible education for responsible AI: a multidisciplinary review and evaluation framework

Maira Klyshbekova , Gisela Reyes Cruz , Caitlin Bentley ,
Stef Garasto , Amy Aisha Brown , Christine Aicardi , Brian Ball ,
Mohammad Naiseh , Oana Andrei

PII: S2666-6596(25)00043-5
DOI: https://doi.org/10.1016/j.jrt.2025.100147
Reference: JRT 100147

To appear in: *Journal of Responsible Technology*

Received date: 7 February 2025
Revised date: 13 December 2025
Accepted date: 26 December 2025

Please cite this article as: Maira Klyshbekova , Gisela Reyes Cruz , Caitlin Bentley , Stef Garasto , Amy Aisha Brown , Christine Aicardi , Brian Ball , Mohammad Naiseh , Oana Andrei , A UK perspective on responsible education for responsible AI: a multidisciplinary review and evaluation framework, *Journal of Responsible Technology* (2025), doi: https://doi.org/10.1016/j.jrt.2025.100147

A UK perspective on responsible education for responsible AI: a multidisciplinary review and evaluation framework

Maira Klyshbekova[a]*, Gisela Reyes Cruz[b], Caitlin Bentley[c], Stef Garasto[d], Amy Aisha Brown[e], Christine Aicardi[f], Brian Ball[g], Mohammad Naiseh[h], Oana Andrei[i].

[a]Department of Digital Humanities, King's College London, London, UK; [b]School of Computer Science, University of Nottingham, Nottingham, UK; [c]Department of Informatics, King's College London, London, UK; [d]School of Computing and Mathematical Sciences, University of Greenwich, London, UK; [e]University of Portsmouth, Portsmouth, UK; [f]Department of Informatics & School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK; [g]Philosophy Faculty, Northeastern University London, London, UK; [h]AFG College, University of Aberdeen, Aberdeen, UK; [i]School of Computing Science, University of Glasgow, Glasgow, UK.

Dr Maira Klyshbekova

mairaklyshbekova@gmail.com *corresponding author

Orcid https://orcid.org/0000-0002-7463-356X

Twitter @MKlyshbekova

Dr Gisela Reyes Cruz

gisela.reyescruz@nottingham.ac.uk

Orcid https://orcid.org/0000-0001-5363-5489

Dr Caitlin Bentley

caitlin.bentley@kcl.ac.uk

Orcid https://orcid.org/0000-0002-2602-601X

Dr Stef Garasto

https://orcid.org/0000-0001-5742-8073

s.garasto@greenwich.ac.uk

Amy Aisha Brown

https://orcid.org/0000-0001-8723-9257

amy_aisha.brown@port.ac.uk

Dr Christine Aicardi

christine.aicardi@kcl.ac.uk

Orcid https://orcid.org/0000-0003-1112-7720

Dr Brian Ball

brian.ball@nulondon.ac.uk

Orcid https://orcid.org/0000-0003-2478-6151

Dr Mohammad Naiseh

mohammad.naiseh@abdn.ac.uk

Orcid http://orcid.org/0000-0002-4927-5086

Dr Oana Andrei

oana.andrei@glasgow.ac.uk

Orcid https://orcid.org/0000-0002-1306-0219

Authors' contributions

MK: Writing – Original draft preparation, Conceptualisation, Investigation, Methodology, Analysis, Writing, Reviewing, Editing.

GRC: Conceptualisation, Investigation, Methodology, Analysis, Writing, Reviewing, Editing.

CB: Supervision, Conceptualisation, Methodology, Analysis, Writing, Reviewing, Editing.

SG: Writing, Reviewing, Editing, Providing feedback.

AAB: Reviewing, Editing, Proofreading, Providing feedback.

CA: Providing feedback, Reviewing, Editing.

BB: Providing feedback, Reviewing, Editing.

MN: Providing feedback, Reviewing.

OA: Providing feedback, Reviewing.

A UK perspective on responsible education for responsible AI: a multidisciplinary review and evaluation framework

Abstract

Responsible Artificial Intelligence (RAI) education has emerged as a way of approaching the field of AI to address a host of concerns (Bentley et al., 2023). Many education providers have been releasing new RAI-related online courses, programmes, or toolkits. When combined with the issues emerging from the development, deployment, and use of AI, the expansion of RAI education and the proliferation of resources raise two critical questions. First, what can we learn about RAI from examining both the content and structure of publicly available RAI educational resources? Second, how might we understand the quality and impact of these RAI resources? We conducted a systematic search of UK RAI educational resources found online. We first present a descriptive analysis of 211 resources collected, including their type, format, cost, sector, audience, and type of provider. Furthermore, we describe our collaborative approach to analysing four pre-selected resources in-depth, from which we outlined an evaluation framework that we then employed for assessing the content of a subset of 47 resources. The five crucial areas of our framework could guide both learners and developers when approaching RAI resources.

Keywords: Responsible AI, Evaluation, Resources, Framework, Multidisciplinary review

1. Introduction

The growing interest in artificial intelligence (AI) has led to numerous negative or unexpected effects (Mikalef et al., 2022). A well-known example is Google's AI image generation tool, Gemini. To address racial bias, Google introduced a technical diversity requirement that produced historically inaccurate images, such as depicting Nazis as people of colour (Shamim, 2024). Public outrage prompted Google to disable the image generation feature, highlighting both the technical and sociopolitical complexities of AI systems. This example illustrates that the complexity of developing and deploying AI responsibly requires many actors to debate and critically assess the socio-technical aspects of AI (Mikalef et al., 2022; Nabavi & Browne, 2023).

AI, as a constellation of technologies and a field of research and practice, consistently eludes definition with its meaning varying widely depending on context, application, and perspective (Sheikh, Prins & Schrijvers, 2023). Likewise, AI is increasingly impacting our societies and environment in both promising and harmful ways, influenced as much by public perception and beliefs as by the capabilities and impacts of AI-enabled systems. This complexity makes it difficult to fully anticipate AI's implications on our lives and societies.

Responsible AI (RAI) has thus emerged as a field of policy, research, and practice concerned with developing, deploying, and governing AI systems in ways that are safe, ethical, and sustainable (Dignum, 2023). RAI education has developed as a means to support diverse people and stakeholders in addressing the complex challenges that AI technologies present (Bentley et al., 2023). Education is an imperative for RAI, with many providers releasing new RAI educational resources (de Laat, 2021; Madaio et al., 2024). However, there remains a noticeable lack of standards or frameworks to evaluate the quality of these resources, leaving little scope to understand their impact or contribution to RAI more broadly.

This gap is particularly problematic given the role that education plays in translating principles into practice. While much current RAI scholarship focuses on the governance of AI in policy and societal levels (Mikalef et al., 2022), studies on how these principles are embedded in the educational resources learners encounter are lacking. This study, therefore, reviewed and evaluated UK-based publicly available RAI educational resources using a tailored multidisciplinary review design, including a scoping review and a multidisciplinary review of selected resources. As part of this review, we examined the characteristics of available resources and whether they support RAI education for various purposes and audiences. However, given the evolving nature of RAI as a field, we also highlighted how multidisciplinary participants assessed the resources. We focused on the UK because of local educational practices and specific conditions, such as sector-specific AI challenges (e.g., NHS struggles with AI bias in healthcare [Nouis et al., 2025]), to offer targeted recommendations for enriching the UK educational landscape. Additionally, as a Responsible Ai UK (RAi UK) research programme funded project, we support the organisation’s mission to promote responsible, inclusive, and publicly beneficial AI education.

Our findings showed a concerning lack of clarity about what constitutes RAI, with unclear learning outcomes and text-heavy materials that focused on basic understanding rather than application. Considering these findings, we present an original evaluation framework to guide learners and developers when approaching RAI resources. The framework includes five key aspects: engaging with the context, building on a socio-technical definition of RAI, questioning the providers' interests, clearly addressing the audience and their needs, and adopting critical pedagogy theory and praxis within RAI education.

The framework addresses a significant gap in the UK's RAI education landscape by focusing on how responsibility is actually learned and taught. It promotes a vision for RAI education that is participatory, critically reflective, and contextually relevant in its practice.

Although this framework emerges from analysing UK-based resources, it could lay the groundwork for developing context-specific frameworks globally, especially given the corporate influence in AI education worldwide.

The paper comprises six sections. The following section provides the literature review on RAI education, its challenges, and the importance of evaluating resources. The next presents the evaluation methodology. The findings are detailed in the fourth section, with the discussion in the fifth. The final section summarises key findings and the framework.

2. Literature review

2.1. Education in Moments of Transition

The term "Responsible AI" has gained widespread attention in both academic and non-academic settings as the adverse impacts of AI systems have become apparent (Mikalef et al., 2022). However, the field remains contested, with ongoing debates about who bears responsibility, what constitutes responsible practice, and how accountability should be distributed across AI systems (Porter et al., 2025). These unresolved tensions risk reducing RAI into a "buzzword", obscuring rather than clarifying the importance of the field (Baeza-Yates, 2023).

The diversity of the RAI field is considerable, frequently adopting incompatible definitions. Dignum (2019) identified RAI as a method for individuals to take responsibility for the power AI brings to them. In contrast, Merhi (2023) argued that RAI concerns establishing standards and values to prevent security, discrimination, and bias issues in AI systems. Yet, these definitions miss the scope and scale of the transformations and actors implicated in RAI design and deployment.

These definitional problems reflect deeper challenges in the field (Reyes-Cruz et al., 2025). As Dignum (2019) observed, RAI is “more than ticking ethical boxes in a report or developing add-on features or switch-off buttons in AI systems” (p. 6). It requires continuous integration of ethical, legal, and societal dimensions into AI design and deployment. This integration involves active participation and engagement of diverse stakeholders (e.g., developers, policymakers, educators, researchers, learners, and the public) who shape systems that are trustworthy in both principle and practice.

Recognising this complexity and the field’s contested status (Reyes-Cruz et al., 2025), we adopt a working definition of RAI as a multidisciplinary field involving policy, research and practice aimed at ensuring development, deployment and governance of AI systems that deliver safe, ethical and sustainable benefits to society. We acknowledge RAI as an evolving field negotiating fundamental challenges related to technology, power, and social responsibility.

Meanwhile, in education, there is an upsurge of interest from international organisations, businesses, governments, and institutions to develop resources that demonstrate their commitment to RAI (Stahl et al., 2023). In fact, as of 2022, there were “over 600 AI-related policy recommendations, guidelines or strategy reports” (Dignum, 2023, p. 196) released by governmental and non-governmental organisations. Large tech companies (e.g., Microsoft, 2023; 2024) are also releasing RAI recommendations and best practices. Higher education institutions (HEIs) are offering training, courses, and degree programmes in RAI to educate various groups. These initiatives vary in nature, scope, and influence as they include product-level improvements, addressing AI-related problems, and principles and guidelines for transforming RAI (Nabavi & Browne, 2023). However, the pressing question is whether these initiatives develop a holistic education (Domínguez Figaredo & Stoyanovich, 2023), encompassing not only the technological aspects but also the societal implications of AI, evaluating them against the field's core aims of ensuring safe, ethical and sustainable AI benefits to society. However, given the complexities within the field and its contested status (Reyes-Cruz et al., 2025), the issue of how to effectively evaluate existing RAI educational resources is both problematic and critically important. The next section draws on RAI and education literature to explore what should be assessed and how.

2.2. Challenges surrounding RAI education and its evaluation

Whilst numerous RAI educational resources exist, there is no research examining how to evaluate these resources or what criteria should guide such assessment. Within education literature, decades of research has churned out valuable insights into pedagogical effectiveness for specific contexts or learners, encompassing schooling, higher education, vocational training, and professional development (Narzulloevna et al., 2020; Senior et al., 2018). However, RAI education presents unique challenges that cross these contexts, as it oversteps traditional disciplinary boundaries and affects virtually every sector of society. Murad (2022) identified three core groups requiring RAI education: civil society, public entities, and system owners. However, these categories mask significant complexity. Within "civil society" alone, which Murad (2022) defined as groups of individuals impacted by AI systems, marginalised groups, and public-interest groups, we find healthcare patients needing to understand data protection, job seekers navigating automated hiring workflows, or citizens affected by algorithmic decision-making in public services. Each group likely needs different levels and types of RAI knowledge and skills. However, it is unclear to what extent diverse needs of varied audiences are presently tailored for in the UK. Additionally, the challenge extends to fundamental questions about learning design. We draw on Biggs and Tang's (2014) concept of constructive alignment because it emphasises examining the relationship between intended learning outcomes, teaching methods, and assessment approaches. For evaluating RAI educational resources, examining this alignment may be particularly important given our focus on whether these resources effectively support learning outcomes that contribute to the aims of RAI.

However, there is very little research on relevant learning outcomes for RAI, let alone learning design. Domínguez Figaredo and Stoyanovich (2023) highlighted that responsibility should be integrated into every stage of the learning process, yet the field lacks frameworks for defining and measuring learning outcomes that balance technical understanding with ethical awareness and practical application. This gap is particularly problematic given the rapid pace of AI technology development and its societal implications.

Likewise, constructive alignment, or the learning design, can be particularly challenging in fields where theory and practice are deeply intertwined (Loughlin et al., 2020). Whilst Lewis and Stoyanovich (2022) advocate for varied materials, such as videos, tests, and simulations, there remains a significant gap in understanding effective learning design of experiences that effectively bridge technical, social, professional, and ethical considerations across different contexts or levels of expertise. Our study addresses these gaps by examining how existing RAI educational resources navigate these challenges, particularly in terms of learning outcome specification, audience targeting, and learning design.

Beyond pedagogical design, evaluating RAI education resources should tackle deeper questions about the field's fundamental challenges related to technology, power, and social responsibility. According to Giroux (1983), education is not merely a means of transmitting knowledge but also a site where dominant social norms and power relations are reproduced through hegemonic processes. In the context of RAI education, this means that educational resources risk reinforcing dominant narratives about AI that could reinforce social injustice or inequality, rather than contributing to societal benefits for all. If RAI education simply teaches compliance with current industry practices without fostering critique of underlying assumptions, it may legitimise, rather than challenge, problematic AI deployments and practices. Freire’s (1970) view on education is crucially important in this context, as education may either facilitate learners to conform to what could be called “irresponsible” AI systems and practices or enable people to “deal critically and creatively with reality and discover how to participate in the transformation of their world” (p. 34). Operationalising this lens involves analysing how resources frame learner agency and responsibility, whose voices or perspectives are included or excluded in the materials, and the types of questioning and dialogue they promote.

Looking beyond the field of education, we also considered relevant RAI and AI ethics frameworks to better analyse how resources frame AI responsibility, whose interests the resources serve, and how learners are envisioned in participating or enacting in RAI development or deployment processes. While several frameworks have been developed to guide ethical design and governance of AI, there is a notable lack of tools to translate these values into educational practices. For instance, Floridi and Cowls (2019) proposed a unified framework of five principles for AI in society: beneficence, non-maleficence, autonomy, justice, and explicability. The framework has the potential to guide the laws and best practices for ethical AI across different social contexts, but it does not address how responsibility can be taught and practised in educational contexts. techUK’s (2025) *recent* report explains the roles, competencies, and pathways needed for operationalising RAI within organisations. While it provides critical aspects for shaping institutional readiness, it does not delve into the pedagogical processes of learning about RAI, which often has to occur "on-the-job" (Madaio et al., 2024). Other traditional quality assurance mechanisms (e.g., Ofsted inspections, QAA benchmarks, or standardised assessments) that rely on fixed curriculum and measurable outcomes are unsuitable for the evolving nature of RAI. Given the significant limitations of existing methods and a lack of tailored evaluation frameworks for RAI education resources, this study addresses these gaps by constructing a novel evaluation framework. As no analytical framework developed from a single disciplinary perspective has been shown to be suitable, we sought out multidisciplinary views and constructed our methodology and evaluation framework in a participatory fashion, as outlined in the next section.

3. Methodology

This study developed a novel evaluation framework through participatory multidisciplinary review. The aims of this study were to 1) identify what makes RAI training effective, and 2) develop approaches for evaluating current RAI offerings in the UK from multiple stakeholder and disciplinary perspectives.

We designed a methodology comprising simultaneous, interconnected activities that allowed our evaluation framework to emerge through: 1) collection of resources, 2) descriptive content analysis of all resources by the lead authors, 3) multidisciplinary curriculum review of four selected resources, and 4) collaborative content analysis of all included resources. This approach enabled us to construct our evaluation criteria through dialogue between diverse disciplinary perspectives while systematically examining the landscape of available resources. This research was granted ethical clearance from King's College London's Research Ethics Committee. In the following sections, we describe each activity.

3.1. Collection of resources

Two authors collected resources to effectively capture the current state of the field. This involved searching and compiling publicly available RAI educational training materials and resources. Thus, our research adopts a scoping review methodology as described by Munn et al. (2018), which is particularly well-suited given the study's aim to identify knowledge gaps and map the available evidence in the chosen field.

The search for resources was conducted using Google from March to July 2024, implementing a systematic keyword search strategy, where we crowdsourced terms from [details blinded for review], leading to 17 variations of "Responsible", 21 variations of "Artificial intelligence", and 13 variations of "Training" (see Table 1).

Our search for RAI educational materials and training yielded a pool of 280 resources documented in a spreadsheet.

Table 1

Variations of Responsible	Variations of Artificial Intelligence	Variations of Training
Responsible	Artificial Intelligence	Training
Ethical	Machine Learning	Course
Fair	Natural Language Processing	Skills
Equity/Gender Equity	Multi-Agent Systems	Learning
Accountable	Intelligent Systems	Education
Human-Centred	Computer Vision	Literacy

Trustworthy	Deep Learning	Resources
Bias Mitigation	Cyber-physical	Materials
Justice	Reinforcement Learning	CPD
Social Justice	Neuro-Symbolic	Executive Education
Sustainable	Conformal Prediction	Toolkit
Rights	Knowledge Graphs	Guidelines
Feminist	Reasoning	Guidance
Inclusive	Planning	
Value Sensitive	Argumentation	
Intersectional	Digital Twins	
Critical	Robotic	
	Control Systems	
	Data	
	Metaverse	
	Augmented Reality	

3.2. Inclusion and exclusion criteria

We included all resources that contained at least one of the keywords from each of our Table 1 columns. We included paid and free resources. We included resources from any type of provider, such as technology and professional services, government, professional training provider, and university. We included any type of format, including online, in-person, text or web-based.

From the list of 280 resources, we excluded five duplicate entries and 64 that were not training or educational resources (13), not available anymore (6), not UK-relevant (26), or did not mention any socio-technical aspect (19). We excluded entries that were not considered training or educational resources, including blog posts and research proposals.

We excluded resources not developed for UK audiences, including only resources in which authors, companies, or providers operated within the UK. We excluded those outside of the UK. In this way, our findings are directly relevant to the UK context, providing a clearer understanding of its gaps and challenges.

We focus on the UK context because this is where we have the most power to influence practice. While the UK is actively advancing AI adoption in multiple sectors, the role of public education and informal learning in building RAI literacy has been underemphasised. There is a clearly demonstrated public demand for high-quality, accessible resources, particularly among those outside formal education (Office for National Statistics, 2023). By focusing on the UK, this research seeks to develop a framework that speaks directly to national education needs, offering tools to support and enhance RAI learning. However, our findings may be relevant to other contexts, which we discuss in the following sections.

Lastly, we excluded resources where the search terms from Table 1 retrieved resources that demonstrated insufficient relevance to RAI. For example, an augmented reality training on business communication skills came up due to the keywords “human-centred” and “augmented reality”, but lacks sufficient relevance to RAI.

3.3. Descriptive content analysis of all resources

Afterwards, a total of 211 resources were coded using descriptive content analysis. Codes included resource ID, coder, name, type of provider, target sector, learning level, audience, organisation, URL, description, learning outcomes, format, cost, duration, and contact details. Some were categorical (e.g., type of provider, target sector, learning level), and others were in a free text format. We predefined some categories but were open to adding more as needed.

3.4. Multidisciplinary curriculum review of four selected resources

We conducted a multidisciplinary curriculum review with 21 participants from UK academic institutions, industry, and civil society who share prior RAI knowledge and experience. The purpose of the review was to evaluate selected resources collaboratively, identifying gaps and areas for improvement. The review also collected information on how participants were evaluating the resources. This multidisciplinary approach was chosen for two reasons. First, AI systems span multiple fields across the sciences and humanities. Second, diverse perspectives are essential for evaluating AI's technical and societal implications.

3.5. Participants

Two multidisciplinary curriculum review workshops were conducted with 21 participants recruited through different channels, such as professional and academic networks and contacts within RAi UK partners, to ensure a diverse range of relevant expertise. Participants represented academia (18), civil society (3), with self-reported knowledge and experience in AI and education. Prior to the workshops, participants provided informed consent and selected their preferred level of confidentiality, including options to remain identified, partially identified, or fully identified. Given the participatory nature of the research and the importance of acknowledging valuable contributions to knowledge co-production, participants were offered the choice to be identified. Two participants remained unidentified, four opted for partial identification, while fifteen chose to be identified. Nine participants contributed to the writing-up process and are listed as co-authors. To protect confidentiality where requested, identifying details have been omitted. Figures 1 to 3 present participants' demographic details such as their positions, disciplines, and affiliations. We include these to show the breadth of disciplines/geographical spread/seniority level. This diversity aimed to capture diverse perspectives on RAI, and by intention, included participants not only from computer science but also from dental education, philosophy, human geography, and other fields. For example, dental education offered insight into healthcare-specific challenges, and philosophy brought critical perspectives on the ethical dimensions of RAI education. This enabled a more nuanced understanding of the technical, social, and educational dimensions that shape current approaches to RAI.

Figure 1. Participants' positions

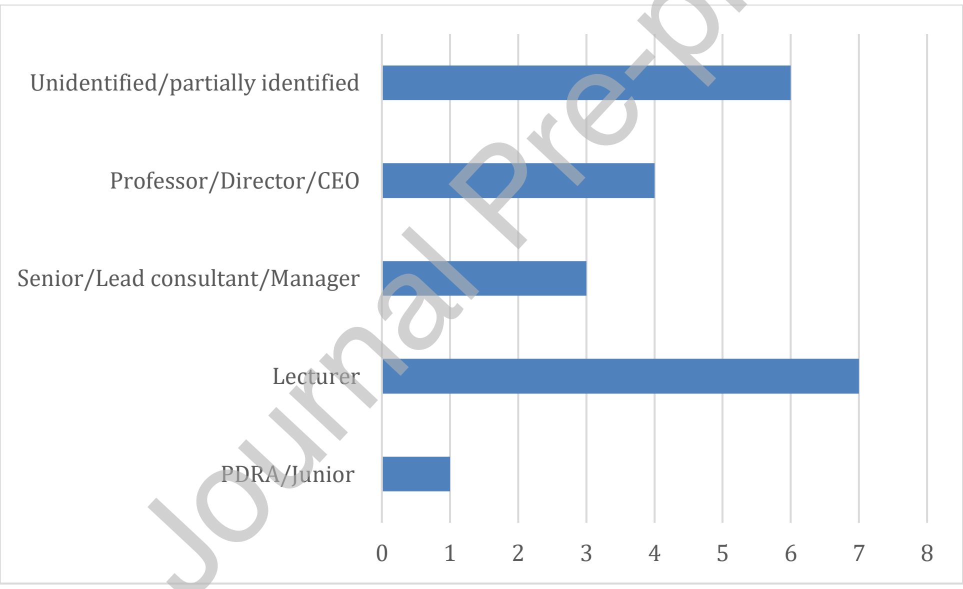


Figure 2. Participants' disciplines

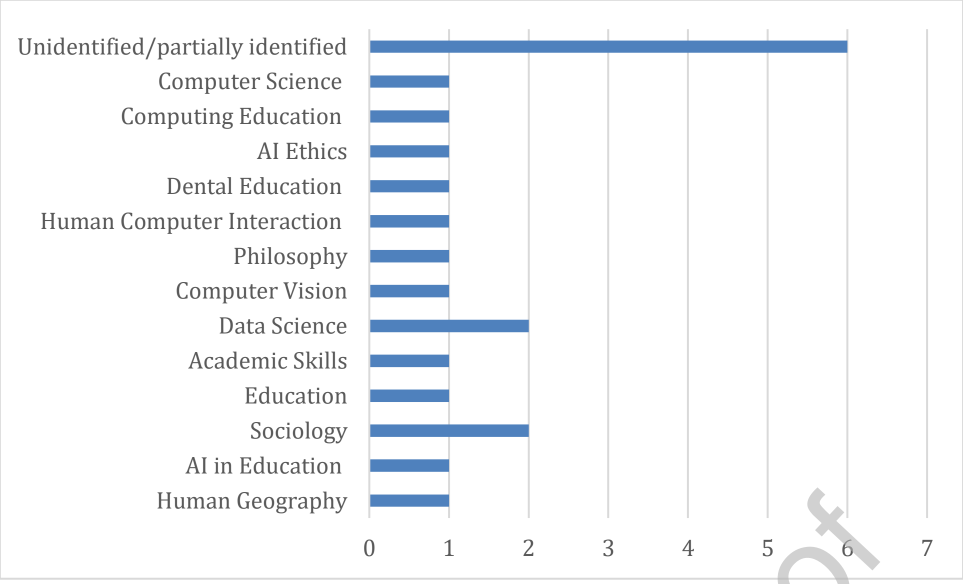
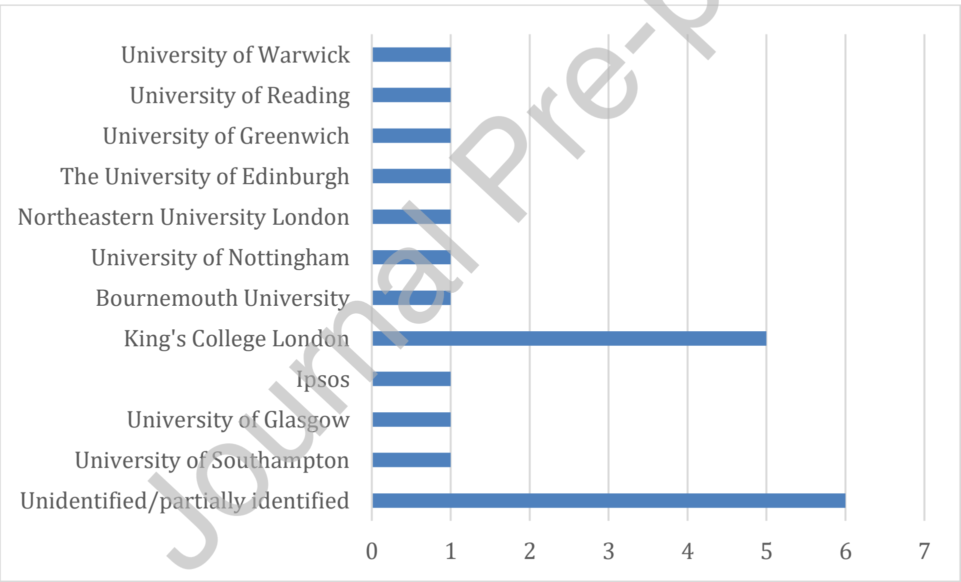


Figure 3. Participants' affiliations



3.6. Multidisciplinary review workshop design

The workshops adopted a participatory knowledge co-production approach where the participants were expected to actively engage in the review activities. Workshops were held online and recorded with participants' informed consent. A maximal variation purposeful sampling technique (Creswell, 2015) was used to select the four resources, which differed in provider type (2 private, 2 public) and audience type (2 general, 2 specific).

Table 2. Resources reviewed in the first workshop

Name	Provider	Workshop	URL
Embrace Responsible AI Principles and Practices	Microsoft (2023)	1	RAI Principles and Practices
Build Responsible Generative AI Application: Introducing the RAFT Framework	Dataiku (2023)	1	The RAFT Framework
Fairness and Responsibility in Human-AI interaction in medical settings	The Alan Turing Institute (2021)	2	Human-AI interaction in medical settings
Generative AI: a problematic illustration of the intersections of racialized gender, race and ethnicity	University of Surrey; University of Glasgow (2024)	2	Gen AI: a problematic illustration

Each workshop began with a brief presentation to communicate the workshop's purpose, engagement ethos, agenda, activities, and selected resources. Participants were divided into two breakout rooms, with each group assigned to review one of the selected resources. The workshops included the following activities:

1. Introductions
2. Familiarising themselves with introductory information of the resource in groups: participants noted their initial impressions in an online collaborative document,

recording the resource description, audience, learning outcomes, learning design, assessment, and creator.

3. Working individually on an assigned section of the resource: participants were asked to scan the resource and then evaluate the specific section assigned to them. They were asked to make notes in the collaborative document, with their evaluation of the section, and the basis for it.
4. Sharing reflections on their section and evaluation of the resource within breakout groups: once each participant finished their evaluation, breakout groups shared their impressions and reflections.

Following the breakout group activities, participants shared their reflections on the evaluation with the wider group, posing the following reflection questions:

1. How does the resource define RAI?
2. How effective is the learning design and why?
3. How could the resource be improved?

We encouraged participants to adopt a critical pedagogy approach when evaluating resources for learning purposes, to uncover how power dynamics, underlying dimensions, and pedagogical approaches that shape RAI are defined and communicated to learners. We maintain that the process of evaluating RAI resources should extend beyond content accuracy and design to assess whether the adopted pedagogical approaches foster critical engagement, participatory learning, and opportunities for reflective dialogue. By making participants aware that they were part of the dialogic process and could act within it, a space was created where they not only shared views but also challenged assumptions, interrogated power dynamics, and explored sociotechnical dimensions of AI.

The workshop ended by inviting participants to contribute to further collaborative research activities.

3.7. Iterative analysis approach

Our analysis process followed three main stages. First, the lead author conducted a thematic analysis on the transcripts from the multidisciplinary review, using an open coding process followed by axial and selective coding. This process resulted in 34 codes, which were merged and collapsed into preliminary categories. The thematic analysis identified the resources' strengths and weaknesses but did not adequately highlight areas for their improvement. Table 3 illustrates how the initial codes were refined and grouped into one of the higher-level themes.

Table 3

Theme	Initial codes	Description	Illustrative excerpt
How RAI defined and conceptualised?	Abstract AI principles, fragmented understanding of RAI, combining RAI and ethics, homogeneous view of AI, understanding of AI	Many definitions for AI principles Heavily emphasised Generative AI Provided definitions were mainly geared towards the end-users or customers rather than the society at large. Responsible AI and Ethical AI as being synonymous, with neither concept having clear definitions. Responsible AI was presented as a way of protecting against harm, rather than as a way of fostering the creation of better products, improving	"One more thing. I think the focus on generative AI too much is bit out of the scope because I think AI is more than just generative AI - it is good to see something about generative AI, but when we talk about Responsible AI, it is more than ChatGPT or any other thing. Especially, I get that example of robotics. Robotics require human or physical interaction and there

		people's lives and making the world a better and fairer place for everyone.	are a lot of risks there that I cannot see here"

In the second stage, to enhance the analytical rigour, we engaged participants in three additional collaborative analysis workshops attended by 9, 12, and 8 participants, respectively.

During the first two workshops, participants familiarised themselves with the data, including both the preliminary categories generated from analysing the transcripts and the online collaborative document notes, providing their feedback on them.

The third workshop concentrated on translating the findings into an evaluation framework based on key questions that emerged from our analysis:

- Does the resource engage with context?
- What are the interests of the provider(s)?
- What are the qualifications of the provider(s)?
- Does the resource address the audience(s) needs?
- What is the level of learner-centric, active, authentic learning?
- Are there concerns about the quality or accuracy of the content?

The third stage focused on applying and evaluating this framework through a detailed analysis of our selected resources. We organised a final workshop to test the evaluation framework by conducting a content analysis. Some questions were slightly revised, and one was made more specific regarding the type of learning and teaching practice participants anticipated (learner-centric, active, authentic learning). However, at this stage, we excluded university-based degree programmes and courses, along with any paid resources, due to access restrictions.

This left 59 resources for in-depth analysis, which were equally distributed among participants for content analysis.

From this process, 12 additional resources were excluded using the same exclusion criteria stated above, yielding a final list of 47 resources evaluated. The findings of that evaluation are presented in the next section.

4. Findings

This section begins with an overview of the collected resources, explaining how RAI training is focused across sectors, highlighting main provider types, target audiences, resource types, costs, and formats. It then presents key themes from the analysis, which revealed three critical questions:

1. How is RAI defined and conceptualised?
2. Who are the target learners and what are the learning outcomes?
3. What are the underlying interests?

4.1. Overview of collected resources

The descriptive coding of the 211 RAI resources initially collected reveals that higher education dominates the UK's RAI training landscape, providing 55% of all coded resources. These are primarily university degree programmes (33%, 70 resources) or courses, including executive education (11%, 23 resources) (Figure 5). Most resources thus require payment (51%, 108 resources) (Figure 7) and are delivered predominantly in person (48%, 102 resources) (Figure 8).

Figure 4. Sector of content

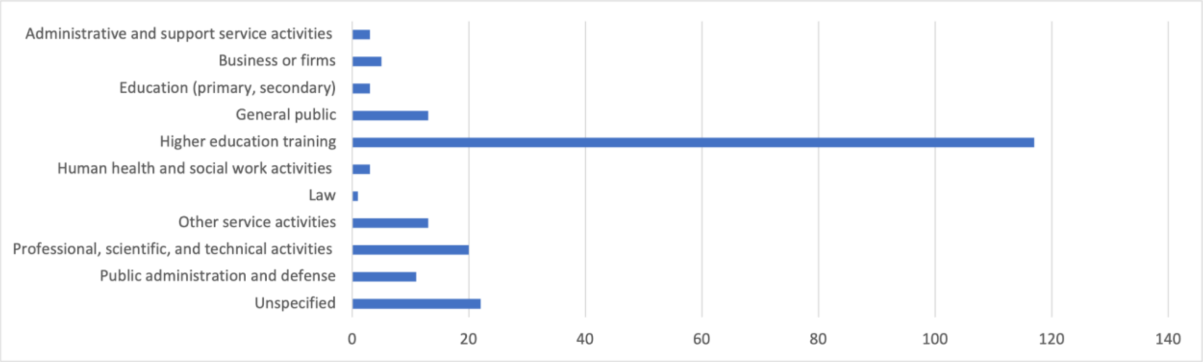


Figure 5. Resource types

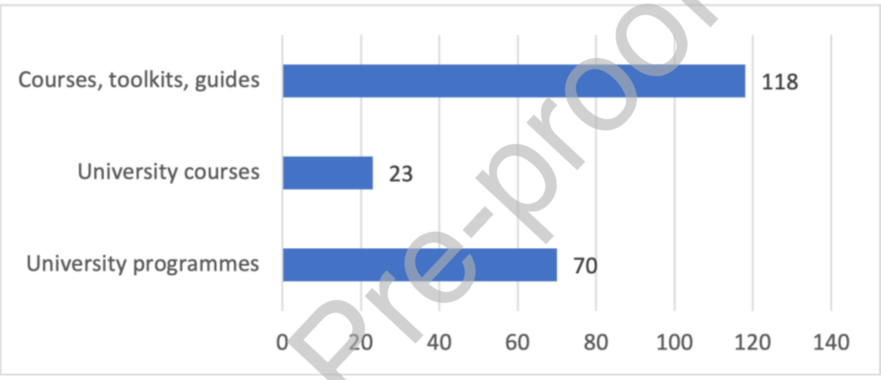


Figure 6. Type of provider

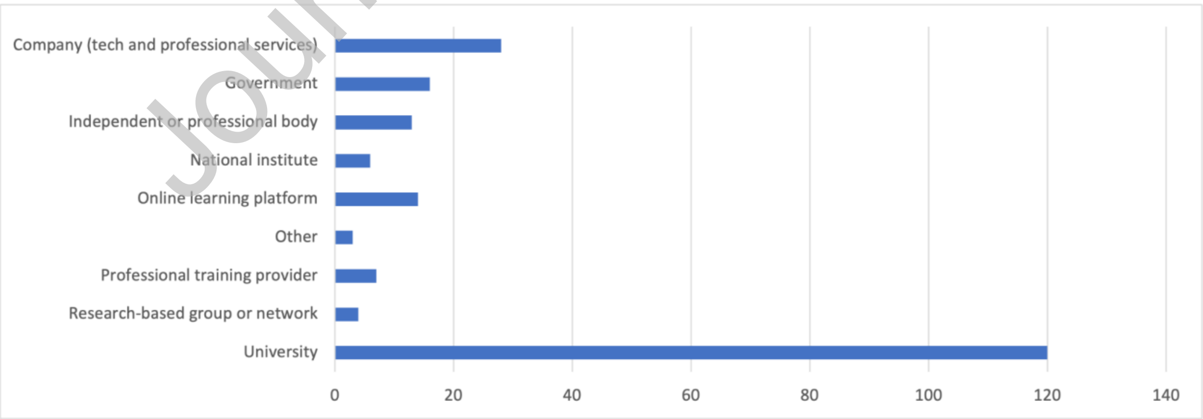


Figure 7. Cost

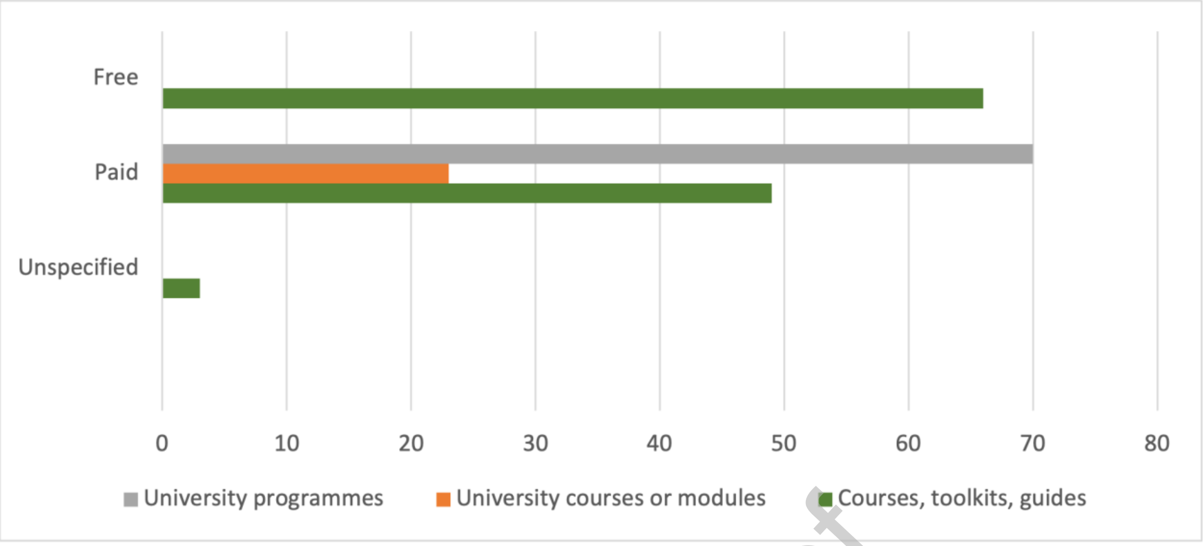
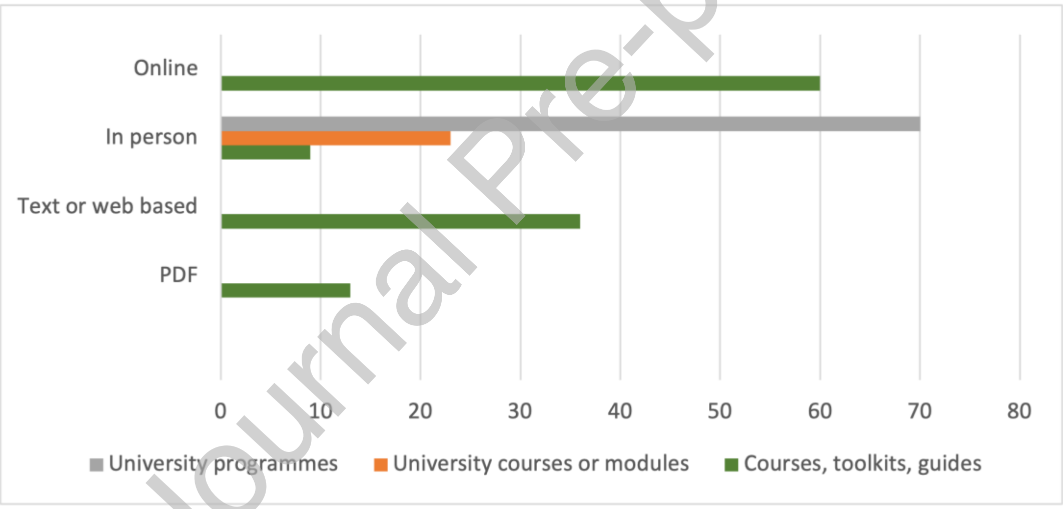


Figure 8. Format



Within higher education, RAI training is concentrated within technical disciplines like Computer Science (39%, 27/70 resources), Engineering (17%, 12/70 resources) and combined Computer Science with Electronics Engineering (14%, 10/70 resources) (Figure 9). However, the extent and approach to RAI education within these programmes is unclear. Moreover, the target audiences for these programmes are students (51%, 108 resources) (Figure 10), but their demographics are unclear.

Figure 9. University Departments

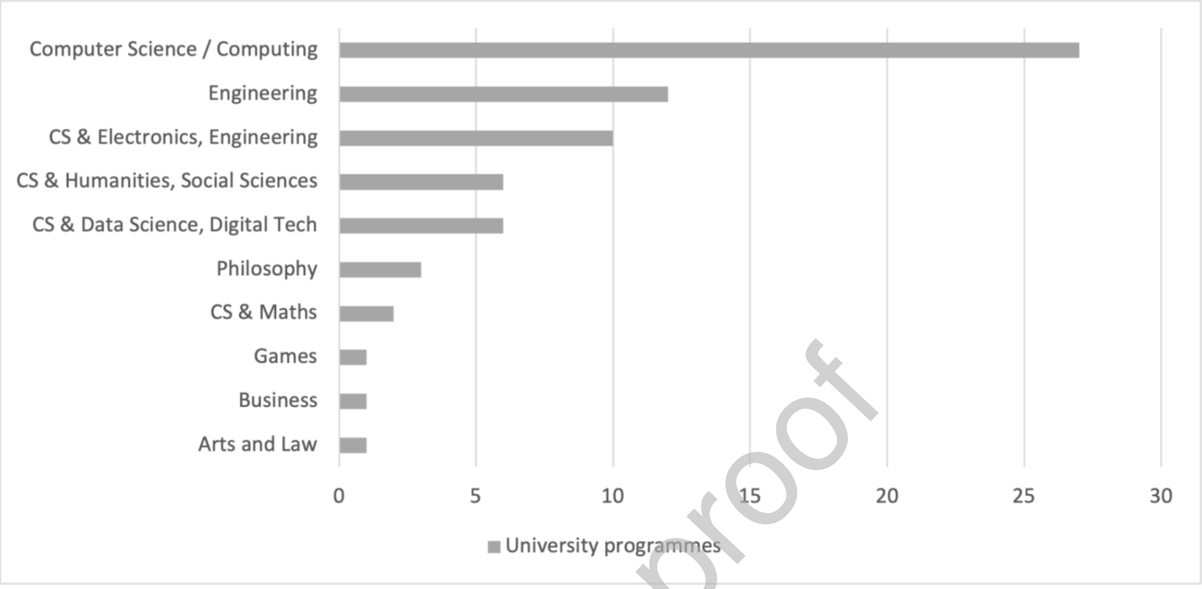
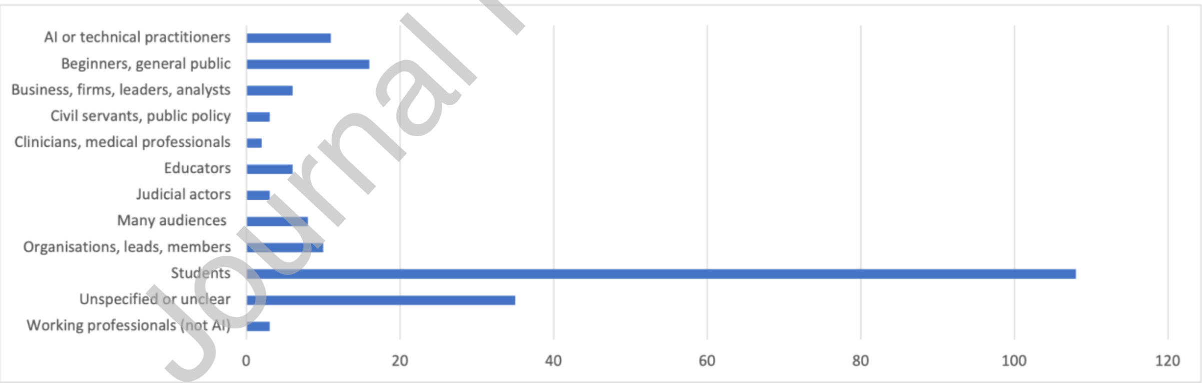


Figure 10. Audience



The non-university sector (green bars in Figure 8) showed different patterns. These resources tended to be delivered online (28%, 60 resources), where about half were short courses (1 hour to 1 day), and half were extended online programmes (2+ days). Other online resources included web content (17%, 36 resources), or PDF guides or toolkits (6%, 13 resources). Unlike university offerings, most of these resources are free (31%, 66 non-university resources), with professional service and technology companies (e.g. Microsoft, PwC) and online learning platforms (e.g. Udemy, DataCamp) being the main providers.

We note two significant gaps from this analysis. First, the lack of audience targeting: 10% (22 resources) have an unspecified sector, and 17% (35 resources) have an unclear target audience. This, combined with the majority representation of students and general public audiences, we find limited provision for key professional groups such as clinicians, educators, or working professionals in specific domains. However, the non-university resources show greater scope and attention to non-technical users and diverse professionals, rather than solely computer scientists or engineers.

Our investigation evaluated the quality and effectiveness of free, short online resources, which represent the main pathway for professional upskilling in RAI. Given that these resources predominantly come from non-university providers operating outside UK academic regulatory frameworks, we aimed to develop an evaluation framework to guide their development and assessment. The following sections, therefore, focus on the results of our analysis of the 47 included resources for in-depth review, combined with the thematic analysis from our multidisciplinary review workshops.

4.2. How is Responsible AI defined and conceptualised?

Our review identified three key aspects that framed the results of our analysis: definition clarity, contextualisation, and societal implications. Most resources (83%) lacked a clear RAI definition, with some citing Microsoft’s six guiding principles of RAI (11%) or focusing on ethics and safety instead (19%). Only 17% of the resources provided an explicit RAI definition. Among those that defined RAI, there was some consistency in approach. For instance, Microsoft (2024) defines it as follows:

> RAI is an approach to developing, assessing, and deploying AI systems in a safe, trustworthy, and ethical way. AI systems are the product of many decisions made by those who develop and deploy them. From system purpose to how people interact with AI systems, RAI can help proactively guide these decisions toward more beneficial and equitable outcomes. (p. 1)

Similarly, another states:

> RAI is a paradigm shift in how we approach AI development and machine learning models. The goal is to create AI that works with ethical considerations safely and fairly. (Synthesia, 2023, p. 1)

These definitions emphasise a safe, fair, and ethical development and deployment of AI. It also touches upon the need to become collectively aware of its benefits and risks and improve its governance based on universal practices. In contrast, some resources took a more nuanced approach:

> There isn't a universal definition of RAI, nor is there a simple checklist or formula that defines how RAI practices should be implemented. Instead, organisations are developing their own AI principles that reflect their own mission and values. (Google Cloud Skills Boost, 2022)

This definition veers away from universal definitions, encouraging flexibility in how organisations develop their own principles that align with their distinct missions and values. However, this type of value-centred rhetoric around RAI seldom specifies whose values are centred, and at which scale they ought to be defined (individual, community, organisation, state, global). Moreover, as Stahl et al. (2022) highlighted, even if there are clear principles and values related to one's mission, in their study, they found it was challenging for experts to agree on how ethical issues could be addressed.

In the multidisciplinary review workshops, none of the four examined resources provided clear RAI definitions. One focused on Generative AI risks without using the term RAI, whilst the other three referenced RAI without defining it. For instance, Microsoft's (2023) resource presented it as a principle-based approach, emphasising principles such as fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability without laying a foundational basis for explaining what RAI is for learners and proceeded to discuss principles and practices, highlighting application over conceptual grounding. Participant 4 expressed their view regarding the resources by stating that:

> I wasn't really persuaded that it was really a good resource for the simple point which is to recite the millions of AI principles that I've seen.

Their perspective aligns with existing research showing that broad AI principles often lack practical implementation guidance (see Kerr et al., 2020).

The second aspect, contextualisation, emerged as crucial precisely because of the limitations we found with principle-based approaches to RAI. Whilst principles were commonly cited in resources, they remained abstract without guidance for implementation in specific settings. For instance, the Microsoft (2023) resource presented principles from Microsoft's point of view on RAI but without a domain or application in a specific context. Similarly, whilst Dataiku's (2023) RAFT framework acknowledged the importance of context by suggesting that businesses consider their unique circumstances, the resource failed to provide concrete examples for different types of organisations. This pattern was widespread, as only 28% of the 47 assessed resources clearly engaged with the context. This lack of contextualisation makes it challenging for learners to understand how RAI principles manifest and should be implemented in their specific domains.

The last aspect that emerged regarding definitional clarity regarded impact. The participants emphasised the need to consider the societal implications of AI, rather than focusing only on organisational perspectives. One of the participants noted, whilst reviewing the Dataiku (2023) resource:

> I kind of feel like before you even can write this guide again you need to decide what is RAI, because it became clear for me when you get to the paper - I was looking at the risks section - in the table on risks there were some good suggestions, but they

> were only for customers. So, there is literally their definition of RAI - it was for their users, and it had no conception of RAI for society at large even though it gives a description of socio-technical. So, I think they have written a guide on how we can be responsible for our users, you know, our customers. And that kind of explains quite a lot as well. Have they even understood what RAI is? (Participant 7)

This limited framing of RAI presents two problems. First, it instrumentalises RAI by reducing it to a tool for organisational risk management rather than a framework for responsible innovation that benefits society. This was evident in how resources approached harm prevention. For instance, at least 31% of the resources mentioned the term “harm” or “harmful”, informing learners of the role of RAI in mitigating harmful consequences. However, they focused narrowly on protecting customers rather than considering broader societal benefits, as Participant 6 stated:

> It's totally geared towards risk management and risk mitigation. So, yes, not at all how to build socially desirable products.

Additionally, this narrow focus fails to acknowledge that AI’s impacts extend far beyond direct users or customers. Participant 4 further explained:

> Because if they really wanted to understand what RAI looked like and articulate it then they would like to go out and talk to the people who would be impacted and have the conversation in their own contexts. Their own framework, the document would be totally different, wouldn’t it? It’ll be like here is how to engage with communities that are impacted by the AI or might be impacted by the AI. I am not sure what practical action anyone would take from the document, which is why I think it needs a rewrite.

Indeed, during discussions between all participants following the resource review, participants supported Participant 4’s argument that it is important not only to understand the technical aspects but also to engage in meaningful conversations with diverse groups to consider the wider societal impact as part of RAI practice.

A more comprehensive conceptualisation was suggested by Participant 10, who proposed RAI as a process incorporating engagement with relevant stakeholders throughout an AI development process:

> Maybe some starting point for RAI education is to think about process of building RAI products. Like what is the process and who is involved in each stage of this process. It will also help to understand what they need to do in each process, who they need to talk to, and how to involve each stakeholder in each process. For instance, at the beginning of building RAI product, how do we define the problem and scope of the AI product, and who should we talk to ensure we're addressing the right issues? When it comes to data collection, how can we ensure it's fair and unbiased, and who should be responsible for auditing this data? Answering these questions will help educators design more comprehensive curricula, allowing them to guide students through the practical application of RAI.

Whilst there may be other ways of developing AI responsibly, this process-oriented view emphasises that RAI education needs to help learners understand not just technical implementation or risk mitigation in immediate business environments but should engage meaningfully with all affected communities. This includes identifying stakeholders beyond immediate users, understanding various contexts of impact, and considering both potential harms and opportunities for societal benefit.

Thus, to clarify what RAI is, participants suggested contextualising RAI and considering its implications for society.

4.3. Who are the target learners and what are the intended learning outcomes?

Our examination of learning outcomes across 47 resources revealed several limitations for equipping learners with foundational knowledge and actionable practices for RAI. First, only 34% of resources reviewed specified learning outcomes, with many failing to balance knowledge acquisition with practical application. For example, the Microsoft (2023) resource provided the following:

- Prepare for the implications of RAI
- Describe principles of RAI
- Establish a system for AI governance
- Take actions for AI governance

- Engage across teams and organizations to implement RAI principles
- Take inspiration from how RAI is approached at Microsoft (p. 1)

Half of these focus primarily on knowledge outcomes (Prepare, Describe, Take inspiration) by describing and interpreting implications and experience. This theoretical emphasis concerned participants, who noted that learners might struggle to translate concepts into their organisational contexts without more specific guidance on application.

The disconnect between theory and practice was further evidenced in resources focused on specific domains. When reviewing The Alan Turing Institute (2021) resource, one of the participants highlighted the lack of clear practical objectives:

> It doesn't really say what the goal is. Like, what's the goal for these clinicians? What are they expected to be able to do differently at the end of this? It's not clear up front. It says things like apply examples and models but seems like there's something missing in terms of what the real goal is here. (Participant 15)

Without clearly defined goals or learning outcomes for practical application, learners may struggle to apply RAI learning in their professional or personal contexts. The second major issue concerned audience specification. Whilst 70% of the 47 resources specified a target audience, many failed to effectively tailor content to their intended learners' backgrounds and needs. This is particularly problematic given the wide impact of AI on society and the need for nuanced understanding across different contexts (Domínguez Figaredo & Stoyanovich, 2023). Moreover, the stage-three reviews showed that most resources targeted "students" in broad terms. For instance, the *Generative AI* resource, which aimed to address AI bias regarding race and gender, illustrated the difficulty with this approach to targeting students without further consideration of student backgrounds. As one participant said:

> Regarding the slide on 'decolonial thinking' there is just too much on that slide. It feels like it's just compressing too many ideas which would put off anybody that was unfamiliar with the literature. There are quite technical phrases that you would need to be in that field to understand; I mean my sense is you need some prior sociological knowledge here. (Participant 17)

This critique relates to a broader issue in RAI education where resources often assume prior knowledge, potentially alienating learners from other disciplinary backgrounds that would benefit from engaging with socio-technical frameworks of RAI.

Similarly, the assumption of AI knowledge and excessive use of technical or business-related vocabulary were also highlighted for posing the risk of disengaging non-technical, non-business audiences who might feel lost and unable to navigate the learning.

The third pedagogical issue identified was the learning design. Returning to the disconnect between theory and practice, participants argued that learning design is critical to enabling learners to make connections between conceptual knowledge and its application. Participants emphasised that the complex nature of RAI requires more interactive and experiential learning design approaches:

> If you think as this as an online course, you know without interaction, you know bouncing ideas, you lose, completely the idea that it's very much like something that is being developed and along the line we have guidelines but we're going to make mistakes and we're going to learn from our mistakes. But there needs to be this reflection process, like along the way. And it's obviously lost if there is not, like this sort of group interaction where everyone bounces ideas like this. (Participant 8)

Likewise, drawing on Kolb’s (1984) experiential learning theory, a cornerstone of adult learning (andragogy), another participant highlighted how the current theory-first approach contradicts established principles of adult education:

> In his model, your starting point is active experimentation from which you reflect on the outcomes, your own feelings and emotions and thoughts. …In contrast, whereas this approach seems to begin with theorising before gets participants to experiment; it’s kind of back to front. (Participant 17)

This reference to Kolb’s (1984) work is particularly relevant for RAI education, as many learners may be adults seeking to update their skills or incorporate RAI practices into their existing roles. Andragogical theory (Knowles, 1978) suggests that such learners benefit most from experiential, reflective approaches that connect to their professional contexts, rather than purely theoretical instruction.

Similarly, the need for flexible paths through the resources reflects this finding as well:

> Having a course that allows for different entry points and routes through it, so doing things like self-assessment at the start, so that people can skip the bits that they don't need to get more information about what they really do need and that kind of thing. I haven't seen any of the AI courses so far offer a kind of multi-entry point way into them and through them. (Participant 15)

In adult education, self-directed learning is important because adults tend to be more motivated by their own personal or professional goals and bring a range of life experiences to learning.

When examining the learning design in the third stage of analysis, we coded 47 resources for the level of learner-centric, active, authentic learning in other resources, highlighting consistent issues across resources. Out of 47 resources, 29 were categorised as “low” by participants, citing reasons such as “No learning outcomes, just a text and five videos. It seems like a collection of different resources”, or “No engagement required on the part of the learner”. The only resource categorised as offering a high level of learner-centredness and engagement included a 2-hour practicum on fairness in machine learning, with articulated learning objectives and hands-on tools to apply the learning.

Overall, this analysis suggests that effective RAI education requires a fundamental rethinking of how resources are designed and delivered, with particular attention to creating learning opportunities that accommodate diverse backgrounds and learning needs.

4.4. What are the underlying interests of resource providers?

Earlier, we showed that HEIs are offering the most education programmes related to RAI. In the UK, HEIs offering formal qualifications are regulated by the Office for Students (OfS) and subject to quality assurance processes through the Quality Assurance Agency (QAA) for higher education. These mechanisms strengthen educational standards and protect learner interests in degree programmes. However, our analysis concentrates on RAI training that falls outside of this regulatory framework, which is crucial because many people lack time to take a year out of their careers to pursue a degree programme. This tendency toward unregulated provision raises important questions about provider qualifications, motivations, and the quality assurance of RAI education.

When reviewing the Dataiku (2023) resource, participants identified characteristics of content marketing rather than educational material:

> I think it's just marketing material. That's my own impression. This is not about informing people about RAI. It's so they can sell their products and along with the product they are bundling this course. And then it's supposed to make you feel safe, but you know you implement their products safely and responsibly but to me, it just feels like marketing material. (Participant 5)

Drawing on professional expertise, another participant reinforced this perspective:

> It's definitely marketing. I've got a marketing background... I feel that they are trying to spread themselves to say this would be useful whatever stage that you are at rather than necessarily specifying other variants which could be industry, sector or, you know, application use, size of organisation, etc. (Participant 7)

This marketing-oriented approach was evident in how resources from tech companies presented RAI concepts – rather than fostering inclusive dialogue or providing practical implementation tools, they often emphasised their own frameworks and products. As Participant 6 noted:

> It was unclear to me whether it was a sales pitch for Microsoft services or an actual training course because whenever it evokes possible topics falling under RAI, there is always a kind of box with a note explaining how Microsoft is doing this very well.

During the review workshops, participants found that resources often presented simplified versions of complex concepts as privacy and data security, whilst prominently featuring the company's own RAI governance approaches. The free availability of these resources, combined with their emphasis on company-specific solutions, suggests they may serve as marketing tools for products and consulting services, rather than learner-centred educational resources.

In the third stage of analysis, we therefore looked specifically for the interests and qualifications of providers, noting any concerns about quality or accuracy of the content. Out of 47 resources, 43% were categorised as having private interest, while 23% showed unclarity regarding who developed the resources and what their qualifications were. This pattern raises concerns about how private interests might be shaping RAI education. Whilst free resources increase accessibility to education, our research shows a need for learners to critically evaluate materials, particularly when they come from private companies, to understand biases and limitations in how RAI concepts are presented and taught. In the next section, we present an evaluation framework to support learners in this vein as our key contribution.

5. Discussion: How should RAI resources be evaluated?

In this research, we set out to examine publicly available RAI educational resources and to assess their quality. Through our review, we found an abundance of RAI educational resources being released by both leading organisations and HEIs. The nature and scope of these resources varied, as they involved in-person and online training, a mix of paid and unpaid materials, courses, and other resources. The critical review of content and learning design showed that: (a) there was a concerning lack of clarity on what constitutes RAI, leaving it to learners to independently interpret and adjust the principles; (b) a lack of clearly defined learning outcomes with text-heavy materials that concentrated mainly on basic understanding rather than application; (c) a clear tendency among resource providers, especially private ones, to prioritise their own interests and goals by leveraging the resource development to promote their products. Given the proliferation of RAI educational resources and their varied approaches, strengths and weaknesses, what is needed is a framework. Thus, informed by our findings, we developed our own evaluation framework consisting of five crucial points that could guide both learners and developers when approaching RAI resources:

- Engage with context
- Build on a socio-technical definition of RAI
- Question the interests of the providers
- Clearly address the audience and their needs
- Adopt critical pedagogy theory and praxis for RAI education.

5.1 Engage with the context

The findings indicated that only half of the reviewed resources seemed to engage with the context and often treat it very broadly. In most cases, the resources did not delve into practical applications or explain contextual implications, making it challenging for learners to apply the knowledge to their specific circumstances. The provided definitions and principles were often discussed in generic, universal terms, overlooking the contextual grounding and the need for an individualistic approach. However, there can be no one-size-fits-all RAI solution (Qiang et al., 2024). Even though it has become a widely used phrase, there is no universal approach that can address and meet the needs of diverse groups. It poses not only a technical challenge but also impacts social and political aspects (Nabavi & Browne, 2023). Often, these interpretations focus on broad concepts such as transparency, ethics, or protection against harm. However, our participants felt that these broad concepts do not provide a nuanced understanding of RAI systems and thus cannot be tailored to specific regions, sectors, and audiences.

Across all themes in our findings, context plays a crucial role in defining what RAI is and how it should operate. This has certain implications for education. In accordance with RRI, which emphasises the need to align technological innovation with broader social values (Casale Mashiah et al., 2023), our findings show that context should be placed at the core and situated in a place, specific industry or professional context. This means that the resources should not only address the technical aspects of RAI but also explore the people involved and the challenges specific to their industries, locations, and applications. As participants suggested, the resources need to involve people and communities that are or might be impacted by AI. By fostering meaningful conversations within specific contexts and communities, the resources will be able to better meet the various needs and challenges that different groups and communities face, and consider their contextual differences.

5.2. Build on a socio-technical definition of RAI

As AI continues to develop at an accelerated pace, it is imperative to adopt a holistic approach to defining RAI. This involves not only focusing on the technical and regulatory aspects of AI but also going beyond them by considering diverse aspects (Bentley et al., 2023). From our review, it was clear that there is a lack of clarity on what constitutes RAI. The resources largely overlooked defining RAI, and when definitions were provided, they tended to reiterate or refer to other known principles. Most importantly, as participants noted, the resources presented a fragmented understanding of RAI by focusing only on its broad concepts and a few themes (e.g., protecting against harm) while overlooking product lifecycle processes. However, RAI is not only about protecting against harm or mitigating risks, but also about improving people’s lives and benefiting society by building better, inclusive products that align with societal values. For the latter, following the precept of engaging with the context, a critical analysis of whose societal values are being prioritised for alignment can also be a key component of RAI development. Therefore, it is important that the resources should build on a holistic definition of RAI, which means defining RAI holistically by incorporating technological, organisational, and societal aspects. Ideally, it should incorporate perspectives from diverse fields such as ethics, law, education, sociology, and technology to ensure that the development of AI is guided not only by technical considerations but also by the needs and values of different stakeholders. In this way, RAI can be a tool that facilitates positive change and fosters trust between different stakeholders, learners, communities, and those who are impacted by AI.

5.3. Question the interests of the providers

As more RAI educational materials are released, it becomes important to critically assess and question the providers' underlying goals. From those collected, it was evident that both private companies and HEIs are actively developing RAI resources. Upon review, participants largely agreed that private companies, especially large tech companies, are leveraging resource development to promote their services and products, heavily pushing their own frameworks of focusing primarily on their areas of expertise without elaborating on others. This made the resources reductive in their topics and pointed to the existing biases surrounding their very purpose. Thus, it becomes crucial for learners to be cautious of these biases and be aware of the providers’ own interests when selecting RAI educational resources. Learners need to understand that the providers are making their resources free and publicly available for a reason.

5.4. Clearly address the audience and their needs

As our findings illustrate, the resources often failed to clearly identify their target audience or indicate which specific group they were designed for. From an educational perspective, it was concerning to find that most of the resources were developed without a specific audience in mind. In the cases when the target audience was identified, they still were too broad, making them difficult to apply to a specific group effectively. As participants noted, the resources that broadly identified the target audience still required some prior knowledge and highlighted the need to narrow the focus for greater relevance. It was proposed to consider learners’ knowledge, backgrounds, and functional roles when developing the resources. This is because an overreliance on technical vocabulary and terms risks disengaging audiences with limited technical knowledge, while oversimplifying content could alienate learners with greater expertise. Thus, the resources should be based on inclusive learning by accommodating various learning needs. While we acknowledge the difficulties of meeting the needs of diverse learners through standalone online resources, RAI educators could consider drawing from the principles of universal design for learning (UDL) to better align with this goal (CAST, 2024). By offering learners various means and formats for engaging with content, resources can better support learners' individual learning needs and preferences. Furthermore, as the resources tended to be tailored towards the users, it is important to highlight that they also need to address other stakeholders, including developers, implementers, and policymakers.

5.5. Adopt critical pedagogy theory and praxis for RAI education

Critical pedagogy for RAI has been explored in different ways. For instance, Goñi et al. (2024) suggested asking critical questions on responsibility, designers and their involvement in it; Bentley et al. (2023) emphasised the significance of contextualisation; and Tahri Sqalli et al. (2023) proposed looking into transparency, fairness and justice, safety and well-being, collaboration, and accountability. We agree with these points, and some of them were also raised in our reviews. However, our review showed that pedagogical theory is another area that appears important for moving forward. Due to concerns about the flooding of the market by private interests, a pedagogical approach that accounts for the regulatory or political influence required to ensure that education benefits all is needed. Other potential learning theories that could inform such pedagogies include Situated Learning theory (see Cobb & Bowers, 1999), which offers a sensible framework to design the resources in a way that integrates learning with real-world application, social interaction, and contextual relevance; this would suggest, for example, an experiential learning type of pedagogy – learning by doing when analysing ethical dilemmas. In another example, a Constructivist Theory approach (see Mills et al., 2006) might lead to a problem-based pedagogical approach. One can spot some different pedagogies in the resources evaluated during the workshops, but those pedagogies were not clearly articulated or implemented.

Moreover, the audience should have the opportunity to reflect on their learning experiences and to question the content of the educational resources. Dialogical approaches (Freire, 1970) could also be employed to critically engage with the materials, moving away from individuals learning in isolation towards groups discussing the topics and learning from one another.

5.6 Novelty of the framework

The framework presented in this study fills a key gap in the current landscape of RAI by moving beyond abstract ethical principles to how responsibility is learned and taught. While existing frameworks, such as Floridi & Cowls (2019) or industry-developed principles like Microsoft’s six guidelines, offer ethical principles for socially beneficial AI, they do not explain how such principles translate into practice or how they should be taught. However, through education, learners develop not only an understanding of AI principles but also learn to interrogate, adapt, and apply these principles in diverse contexts. Other policy-oriented efforts, such as techUK’s (2025) report, focus on standardising RAI through governance and professional roles, but do not address pedagogical concerns. In contrast, the proposed framework emphasises education as a vector of influence on RAI itself and foregrounds such critical questions as what is RAI? Who decides this? How are learners educated, by whom, and in what contexts? By examining resources according to the five dimensions of evaluation, the framework reinforces RAI education as a more participatory, critically reflective, and contextually relevant practice.

However, we acknowledge that this study, while insightful, is subject to contextual and methodological limitations. First, our scope was deliberately limited to UK-based resources. While this enabled a more focused contextual analysis, it may limit the generalisability of our conclusions to other contexts. Second, our focus on freely accessible online resources meant that institutionally embedded courses or subscription-based resources were excluded. Future research could expand the scope and refine the selection criteria. Despite these limitations, the framework has practical application. In the UK context, RAi UK is using it to inform the evaluation of RAI resources and materials to be hosted on its platform, highlighting its real-world applicability.

6. Conclusion

Our descriptive analysis findings showed that HEIs, along with technology and professional service companies, are leading the development of RAI educational resources. HEI courses and programmes are often dominated by Computer Science and Engineering departments. Other educational resources not provided by HEIs take the form of online and in-person courses, text, web-based, or PDF materials.

Our content analysis raised multiple concerns in the resources we had access to, such as lack of clarity about what RAI entails, who the intended audiences are, leveraging of the resources to promote the providers' own products and services, and the provision of introductory concepts without equipping the learners on how to apply the knowledge in practice.

This article and the framework presented in the previous section aim to guide resource developers and learners in making informed decisions about RAI education. Building on the core aspects identified in the review of RAI educational resources, our framework provides a structured pathway for developers and learners to navigate the complex landscape of RAI.

Funding

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from Responsible Ai UK.

Acknowledgements

We thank all the participants for their contribution and their willingness to share their thoughts and insights.

References

The Alan Turing Institute. (2021). Fairness and Responsibility in Human-AI interaction in medical settings. Retrieved March 6, 2024 06.03.2024 https://www.turing.ac.uk/courses/fairness-and-responsibility-human-ai-interaction-medical-settings

Baeza-Yates, R. (2023). Lecture held at the Academia Europaea Building Bridges Conference 2022: An introduction to responsible AI. *European Review, 31*(4), 406–421. https://doi.org/10.1017/S1062798723000145

Bentley, C., Aicardi, C., Poveda, S., Magela Cunha, L., Kohan Marzagao, D., Glover, R., Rigley, E., Walker, S., Compton, M., & Acar, O. A. (2023). *A framework for responsible AI education: A working paper*. SSRN Working Paper Series. http://dx.doi.org/10.2139/ssrn.4544010

Biggs, J., & Tang, C. (2014). Constructive alignment: An outcomes-based approach to teaching anatomy. In L. K. Chan & W. Pawlina (Eds.), *Teaching anatomy: A practical guide* (pp. 31–38). Springer International Publishing. https://doi.org/10.1007/978-3-319-08930-0_4

Casale Mashiah, D., Beeri, I., Vigoda-Gadot, E., & Hartman, A. (2023). Responsible research and innovation in Europe: Empirical evidence from regional planning initiatives in Austria, Norway, and Spain. *European Planning Studies*, *31*(9), 1949–1974. https://doi.org/10.1080/09654313.2023.2170215

CAST. (2024). *Universal Design for Learning Guidelines version 3.0.* https://udlguidelines.cast.org

Cobb, P., & Bowers, J. (1999). Cognitive and situated learning perspectives in theory and practice. *Educational researcher*, *28*(2), 4–15.https://doi.org/10.3102/0013189X028002004

Creswell, J. W. (2015). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Pearson.

Dataiku. (2023). *Build Responsible Generative AI Application: Introducing the RAFT Framework*. https://web.archive.org/web/20240302234007/https://pages.dataiku.com/responsible-generative-ai

de Laat, P. B. (2021). Companies committed to responsible AI: From principles towards implementation and regulation? *Philosophy & technology*, *34*(4), 1135–1193. https://doi.org/10.1007/s13347-021-00474-3

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer. https://doi.org/10.1007/978-3-030-30371-6

Dignum, V. (2023). Responsible Artificial Intelligence---From Principles to Practice: A Keynote at TheWebConf 2022. *ACM SIGIR Forum 56*(1), 1–6. ACM. https://doi.org/10.1145/3582524.3582529

Domínguez Figaredo, D., & Stoyanovich, J. (2023). Responsible AI literacy: A stakeholder-first approach. *Big Data & Society*, *10*(2). https://doi.org/10.1177/20539517231219958

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Freire, P. (1970). *Pedagogy of the oppressed*. Continuum Books.

Giroux, H. A. (1983). *Theory and Resistance in Education: A Pedagogy for the Opposition*. Bergin & Garvey.

Goñi, J. “Iñaki,” Rodrigues, E., Parga, M. J., Illanes, M., & Millán, M. J. (2024). Tooling with ethics in technology: A scoping review of responsible research and innovation tools. *Journal of Responsible Innovation*, *11*(1). https://doi.org/10.1080/23299460.2024.2360228

Google Cloud Skills Boost. (2022). *Introduction to Responsible AI*. Retrieved June 20, 2024, from https://www.cloudskillsboost.google/course_templates/554?_gl=1*121yiqr*_up*MQ..*_ga*

NTczODU5NzQ4LjE3MzczOTczNjQ.*_ga_2X30ZRBDSG*MTczNzM5NzM2My4xLjAuMTczNzM5NzM2My4wLjAuMA

Kerr, A., Barry, M., & Kelleher, J. D. (2020). Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951720915939

Knowles, M. S. (1978). Andragogy: Adult learning theory in perspective. *Community College Review*, *5*(3), 9–20. https://doi.org/10.1177/009155217800500302

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Prentice-Hall.

Lewis, A., & Stoyanovich, J. (2022). Teaching responsible data science: Charting new pedagogical territory. *International Journal of Artificial Intelligence in Education*, 1─25. https://doi.org/10.1007/s40593-021-00241-7

Loughlin, C., Lygo-Baker, S., & Lindberg-Sand, Å. (2020). Reclaiming constructive alignment. *European Journal of Higher Education*, *11*(2), 119–136. https://doi.org/10.1080/21568235.2020.1816197

Madaio, M., Kapania, S., Qadri, R., Wang, D., Zaldivar, A., Denton, R., & Wilcox, L. (2024, June). Learning about Responsible AI On-The-Job: Learning Pathways, Orientations, and Aspirations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1544–1558).

Merhi, M. I. (2023). An assessment of the barriers impacting responsible artificial intelligence. *Information Systems Frontiers, 25*(3), 1147–1160. https://doi.org/10.1007/s10796-022-10276-3

Microsoft. (2023). *Embrace Responsible AI Principles and Practices*. Retrieved March 5, 2024, from https://learn.microsoft.com/en-us/training/modules/embrace-responsible-ai-principles-practices/

Microsoft. (2024). *What is Responsible AI?*. https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai?view=azureml-api-2

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems*, *31*(3), 257–268. https://doi.org/10.1080/0960085X.2022.2026621

Mills, J., Bonner, A., & Francis, K. (2006). The development of constructivist grounded theory. *International Journal of Qualitative Methods*, *5*(1), 25–35. https://doi.org/10.1177/160940690600500103

Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, *18*, 1–7. https://doi.org/10.1186/s12874-018-0611-x

Murad, M. (2022). Back to basics: Revisiting the responsible AI framework. *Towards Data Science*. https://towardsdatascience.com/back-to-basics-revisiting-the-responsible-ai-framework-847fd3ec860b

Nabavi, E., & Browne, C. (2023). Leverage zones in Responsible AI: Towards a systems thinking conceptualization. *Humanities and Social Sciences Communications*, *10*(1), 1–9. https://doi.org/10.1057/s41599-023-01579-0

Narzulloevna, A. S., Tokhirovna, K. F., Bakhramovna, A. Z., Odeldjanovna, S. S., & Jurayevich, S. U. (2020). Modern pedagogical methods in effective organization of lessons. *Journal of Critical Reviews*, *7*(9), 129–133. http://dx.doi.org/10.31838/jcr.07.09.24

Nouis, S. C., Uren, V., & Jariwala, S. (2025). Evaluating accountability, transparency, and bias in AI-assisted healthcare decision-making: a qualitative study of healthcare professionals' perspectives in the UK. *BMC Medical Ethics*, *26*(1), 89. https://doi.org/10.1186/s12910-025-01243-z

Office for National Statistics. (2023). *Public awareness, opinions and expectations about artificial intelligence: July to October 2023.* https://www.ons.gov.uk/businessindustryandtrade/itandinternetindustry/articles/publicawarenessopinionsandexpectationsaboutartificialintelligence/julytoctober2023

Porter, Z., Ryan, P., Morgan, P., Al-Qaddoumi, J., Twomey, B., Noordhof, P., McDermid, J., & Habli, I. (2025). *Unravelling responsibility for AI. Journal of Responsible Technology, 100124*. https://doi.org/10.1016/j.jrt.2025.100124

Qiang, V., Rhim, J., & Moon, A. (2024). No such thing as one-size-fits-all in AI ethics frameworks: A comparative case study. *AI & Society*, *39*(4), 1975–1994. https://doi.org/10.1007/s00146-023-01653-w

Reyes-Cruz, G., Perez Vallejos, E., Barnard, P., Schneiders, E., Tachiquin, M., Price, D., Eke, D., Dowthwaite, L., Gomez Bergin, A., Portillo, V., & Fischer, J. (2025). Unravelling Responsible AI: An umbrella review. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8*(3), 2177–2188. https://doi.org/10.1609/aies.v8i3.36704

Senior, C., Fung, D., Howard, C., & Senior, R. (2018). What is the role for effective pedagogy in contemporary higher education? *Frontiers in Psychology, 9*, 1299. https://doi.org/10.3389/fpsyg.2018.01299

Shamim, S. (2024). Why Google's AI tool was slammed for showing images of people of colour. *Al Jazeera*. https://www.aljazeera.com/news/2024/3/9/why-google-gemini-wont-show-you-white-people

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial intelligence: definition and background. In *Mission AI: The new system technology* (pp. 15–41). Springer International Publishing. https://doi.org/10.1007/978-3-031-21448-6_2

Stahl, B. C., Antoniou, J., Bhalla, N., Brooks, L., Jansen, P., Lindqvist, B., Kirichenko, A., Marchal, S., Rodrigues, R., Santiago, N., Warso, Z., & Wright, D. (2023). A systematic review of artificial intelligence impact assessments. *Artificial Intelligence Review 56*(11), 12799–12831. https://doi.org/10.1007/s10462-023-10420-8

Stahl, B. C., Antoniou, J., Ryan, M., Macnish, K., & Jiya, T. (2022). Organisational responses to the ethical issues of artificial intelligence. *AI & Society*, *37*(1), 23–37. https://doi.org/10.1007/s00146-021-01148-6

Synthesia. (2023). *Responsible AI: Your in-depth guide to safe AI.* https://www.synthesia.io/post/responsible-ai

Tahri Sqalli, M., Aslonov, B., Gafurov, M., & Nurmatov, S. (2023). Humanizing AI in medical training: Ethical framework for responsible design. *Frontiers in Artificial Intelligence*, *6*, 1189914. https://doi.org/10.3389/frai.2023.1189914

techUK. *(2025) Mapping the responsible AI profession, a field in formation. https://www.techuk.org/resource/techuk-paper-mapping-the-responsible-ai-profession-a-field-in-formation.html*

Declaration of Interest Statement

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for this journal and was not involved in the editorial review or the decision to publish this article.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: